NL021204
dec

PCT

WORLD INTELLECTUAL PROPERTY ORGANIZATION
International Bureau

INTERNATIONAL APPLICATION PUBLISHED UNDER THE PATENT COOPERATION TREATY (PCT)

(51) International Patent Classification 6 :
G06F 13/00, H04N 7/173, H04L 12/54

A2

(11) International Publication Number: WO 97/48049

(43) International Publication Date: 18 December 1997 (18.12.97)

(21) International Application Number: PCT/FI97/00366

(22) International Filing Date: 11 June 1997 (11.06.97)

(30) Priority Data:
962467 13 June 1996 (13.06.96) FI

(81) Designated States: CA, JP, US, European patent (AT, BE, CH, DE, DK, ES, FI, FR, GB, GR, IE, IT, LU, MC, NL, PT, SE).

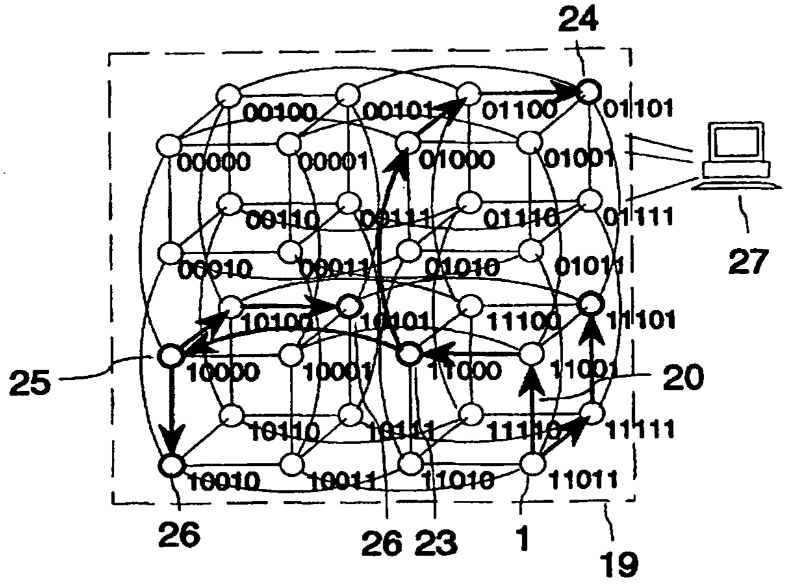
Published
In English translation (filed in Finnish).
Without international search report and to be republished upon receipt of that report.

(71)(72) Applicant and Inventor: OLLIKAINEN, Ville, Juhana [FI/FI]; Kämnerinkuja 2 D 19, FIN-00750 Helsinki (FI).

(54) Title: FILE SERVER WITH A CONFIGURATION SUITED FOR DISTRIBUTION OF DECENTRALIZED DATA

(57) Abstract

The present invention is related to a file server particularly suited for video-on-demand service based on decentralization of files contained in the nodes over a configuration formed by nodes serving a single user or multiple users, each equipped with a magnetic or optical mass memory unit. Such nodes are designed to serve not only their own users but also other nodes of the decentralized file server so that files incoming to the node can be distributed forward to at least two other nodes of the server. A file fetched by several users simultaneously is copied automatically into a plurality of nodes in the file server. This arrangement permits the data transfer rate required from the magnetic or optical memory units to be defined as a constant which is independent from the number of users, mode of use and file server size without any compromise in the performance of the file server.



FOR THE PURPOSES OF INFORMATION ONLY

Codes used to identify States party to the PCT on the front pages of pamphlets publishing international applications under the PCT.

AL	Albania	ES	Spain	LS	Lesotho	SI	Slovenia
AM	Armenia	FI	Finland	LT	Lithuania	SK	Slovakia
AT	Austria	FR	France	LU	Luxembourg	SN	Senegal
AU	Australia	GA	Gabon	LV	Latvia	SZ	Swaziland
AZ	Azerbaijan	GB	United Kingdom	MC	Monaco	TD	Chad
BA	Bosnia and Herzegovina	GE	Georgia	MD	Republic of Moldova	TG	Togo
BB	Barbados	GH	Ghana	MG	Madagascar	TJ	Tajikistan
BE	Belgium	GN	Guinea	MK	The former Yugoslav Republic of Macedonia	TM	Turkmenistan
BF	Burkina Faso	GR	Greece			TR	Turkey
BG	Bulgaria	HU	Hungary	ML	Mali	TT	Trinidad and Tobago
BJ	Benin	IE	Ireland	MN	Mongolia	UA	Ukraine
BR	Brazil	IL	Israel	MR	Mauritania	UG	Uganda
BY	Belarus	IS	Iceland	MW	Malawi	US	United States of America
CA	Canada	IT	Italy	MX	Mexico	UZ	Uzbekistan
CF	Central African Republic	JP	Japan	NE	Niger	VN	Viet Nam
CG	Congo	KE	Kenya	NL	Netherlands	YU	Yugoslavia
CH	Switzerland	KG	Kyrgyzstan	NO	Norway	ZW	Zimbabwe
CI	Côte d'Ivoire	KP	Democratic People's Republic of Korea	NZ	New Zealand		
CM	Cameroon			PL	Poland		
CN	China	KR	Republic of Korea	PT	Portugal		
CU	Cuba	KZ	Kazakstan	RO	Romania		
CZ	Czech Republic	LC	Saint Lucia	RU	Russian Federation		
DE	Germany	LI	Liechtenstein	SD	Sudan		
DK	Denmark	LK	Sri Lanka	SE	Sweden		
EE	Estonia	LR	Liberia	SG	Singapore		

File server with a configuration suited for distribution of decentralized data

The present invention relates to a file server comprised of nodes with means for storage of decentralized data and means for emission of user-requested information, said server being characterized by an uncomplicated structure and almost unlimited expandability. Additionally, the server is capable of acting as a high-capacity, fast and fault-tolerant buffer memory between users and information suppliers.

The invention is primarily intended to provide a sustained data transfer to a plurality of simultaneous users. Hence, the invention is particularly suited for use in video-on-demand systems in which a single server may provide services up to thousands of users. Approaches based on conventional server architectures stumble on problems caused by limitations in data transfer rates and seek times of storage devices, file fragmentation, data management and unsatisfactory fault tolerance, whereby any attempts to solve these problems result in an appreciable rise in the cost of servers.

Several patents and papers have been disclosed in the art of parallel processing and file server architectures. An example of a video-on-demand server based on a segmented video on-demand system is described in the European patent application EP0633694A1. The hypercube architecture known in the art as a typical multinode message-passing server configuration was introduced in the mid-70's (cf. W. Millard: Hyperdimensional Microprocessor Collection Seen Functioning as Mainframe, Digital Design, 1975, Vol. 5, No. 11, p. 20), and later the hypercube has established its position as a widely known parallel-processing architecture. A hierarchical network resembling a hypercube is disclosed in US Pat. No. 5,471,580.

The present invention is based on distributing the user-requested files over the nodes 1 serving at least one user, whereby the nodes can not only serve their own users but also distribute files over data transfer links 20 to other nodes 23 with a relatively low latency. The files requested by several users simultaneously are copied in a chain-reaction-like fashion to all the nodes forming the server.

Simultaneously, an improved fault tolerance of the server is attained, because a server consisting of several nodes may now be inexpensively constructed with the additional advantage that no single failure can stop the function of the entire server system.

Nodes 1 defined in detail in appended claims 1 - 6 are connected together to form a file server defined in appended claims 7 - 12; and as preferred embodiment of the invention, next will be described a file server with a structure of an N-dimensional hypercube 19 based on the nodes defined in appended claim 11 and the server configuration defined in appended claim 12.

The structure of a typical hypercube 19 and one approach to the identification of the nodes 1 is presented in Figs. 3A-3F. In the beginning, the hypercube is 0-dimensional and it has only one node 1. Next, in a 1-dimensional hypercube there is also a copy of the original node, and a data transfer link 20 exists between these two nodes. The original node is identified by number 0 and the copy by number 1. An N-dimensional hypercube can be formed by copying a N-1 dimensional template hypercube and connecting the corresponding nodes with data transfer links. In the hypercube copying process, in front of the addresses of the nodes of the template hypercube is added one bit which has a value 0 for the node addresses of the template hypercube and a value 1 for the addresses of copied hypercube. By virtue of this numbering scheme, the binary address of any two adjacent nodes differ by one bit only.

Accordingly, two adjacent nodes 1 are always connected by a data transfer link 20. The node with specifications defined in appended claim 1 is able to transparently route data transfers between other nodes of the server. The distance d between two communicating nodes is defined as the smallest number of data transfer links 20 needed for data communication between these nodes. When the nodes 1 are arranged in the form of hypercube 19 as defined in appended claim 12 with the addresses of the nodes 1 numbered as in Figs. 3A-3F (whereby the binary addresses of any two adjacent nodes can differ by one bit only), the distance between two nodes is obtained by applying the exclusive-OR (XOR) operator to these binary numbers and counting the number of 1's in the result. Then, the amount of shortest possible data transfer routes is the factorial d! of the distance. Assuming that every data transfer link has an equal probability of being already in use, the probability of finding a free route between the transmitting 23 and receiving 24 nodes will be the higher the greater the distance between these nodes. This means that user accessibility to the files will be the better the larger the user base for which the server is constructed.

In conventional server architectures, constraints set by the data storage seek time and data transfer rate are the most common bottlenecks limiting the expandability of the server. These problems can be avoided by limiting the number of users 5 and simultaneous internode data transfers 6, yet availing the inexpensive technology of the present invention.

In practice, the server operates as follows (refer to Fig. 4). When a user requests a file, his node 25 contacts the central management system 27 that checks which one(s) of the nodes 1, 23, 24 have the requested file. Next, a route is established from the closest routeable node 23 to the user's node and the data transfer is initiated. Immediately at the beginning of the data transfer, the central management system 27 registers the requested file being available also from the user's node 25. Resultingly, files requested simultaneously by a plurality of users will be simultaneously available in real time from a plurality of nodes 1, 23, 24, 25 to new nodes 26 requesting the file in question.

The data communication nodes 21, 22 defined in appended claim 6 are able to transfer files, e.g., from external servers or from a high-capacity data storage device.

As described in appended claim 10, there can be several data communication nodes with the same data contents in a single server system. The user-requested file may be fetched from the data communication nodes if the file data is not available from any user nodes at a reasonably short distance. If the system detects that none of the user nodes contains a minimum amount of file data sufficient for assuring start of sustained data transfer from these external sources, the file must be copied to the data storage 2 of the data communication node prior to passing it to the internal data transfer links 20.

The node 1 has three different functions: it operates as a routing channel 3 between the other nodes, provides means of data transfer from the server to its users and stores data in its memory for further distribution.

As defined in claim 3, an elastic ring buffer built into the data storage means 2 of the node operates as follows.

First, all the files requested by the user are written to the elastic ring buffer 13. In the ring buffer there is one write pointer 15, which serves the storage of incoming data 17 into the buffer, and several read pointers 14. The maximum number of read pointers is limited by the seek time and data transfer rate of the data storage means 2. The read pointers 14 serve the user of the ring buffer 13 and those nodes 1 which are copying files from the ring buffer 13 into their own data storage means 2. When the write pointer 15 passes by the beginning of any file 18, this file is marked and registered by the central management system 27 as no longer available from this ring buffer 13.

Use of the elastic ring buffer 13 offers significant advantages: it makes real-time data management easier in the server while simultaneously giving the user a constant amount of storage capacity which is in an uncomplicated manner available to the user. As an additional benefit, data fragmentation in the ring buffer is practically nonexistent. The buffer also makes file management easier: statistically, the life span of a file in the ring buffer will be longer if the file is in frequent use over an extended period of time, while files subject to random and rare requests will have only a short life span. When necessary, the central management unit 27 can have any file copied into a storage area outside the ring buffers, either into the storage means 2 of a node, or into a high-capacity magnetic or optical storage of the data transfer node defined in claim 6.

As defined in claim 7, transfer of a file into the storage means 2 of a node takes always place at a rate faster than real time. Consequently, a file 17 is always playable and available from the storage means 2 immediately after the data transfer has been initiated. For the same reason, all the files 16 having their starting point copied into the ring buffer 13 are readable from there, from the beginning to the end. While the above description and the diagram of Fig. 2 refer to the use of a ring buffer, data transfer at a rate faster than real time can assure even without the ring buffer a faster release of the memory space than is required for the write operation; however, the management of data transfer in a node not equipped with a ring buffer 13 defined in claim 3 becomes essentially more complicated.

Slowing down of the write pointer 15 as defined in claim 3 whenever it reaches any read pointer 14 guarantees undisturbed read for any file. Analogous, the faster than real time movement of the read pointer 14 defined in claim 7 assures that the read pointer moves aside faster than required for undisturbed read of any incoming file 17.

The above-described technique represents an exemplifying embodiment of the implementation of a file server by virtue of using nodes defined in claims 1, 2, 3, 4 or 5. According to claim 8, a still higher bit rate can be achieved by clustering several nodes 1 to serve a single user, e.g., by synchronizing the outputs of and combining the data from these nodes into a single bit stream. It should also be noted, that node according to claim 1 makes it also possible to implement multiple parallel data connections into the data channels serving as the data transfer links between the nodes using, e.g., any conventional method of modulation, alternation or contention.

All the functions of the central management system 27 can be decentralized to the nodes 1 which in such a configuration are arranged to communicate with each other and establish routes without any separate controller. This is possible, because the central management system 27 does not have to contain any information which would not be as readily available from the nodes 1.

The use of a hypercube architecture as defined in claim 12 not only gives an easy-to-manage symmetrical structure and the possibility of using identical nodes, but also offers a server availability which does not decrease when the size of the server increases.

The elucidate this fact, a situation will be contemplated in which only one copy of each file has been made, each copy into a separate node, and every data transfer link carries only one data transfer at a time. This state may be considered to represent a worst case situation, because routing inside the server will become easier after several copies of a file in different nodes are available. Then, an average fetch distance for the data transfer of a file is N/2, reserving N/2 data transfer links 20 between the nodes 1. Related to this case, three important findings are made:

1) Inasmuch as the server has $(N/2)*2^N$ data transfer links and 2^N users, the number of data transfer links increases at the same rate as the need for the links. Resultingly, the probability of any particular link being already reserved remains constant.

2) As an average, $(N/2)!$ different routes are available for a data transfer, and moreover, the probability of all of them being reserved simultaneously decreases when N increases.

3) In a larger multinode server, the probability of finding a multiple copies of a requested file is also higher. This means that the search for a free route can be extended to a greater number of nodes and probability of finding a free route to a desired node increases.

Claims:

1. A node-like data transfer facility accessible to at least one user , later called a node (1), intended for the distribution of files, particularly video and audio files, said node comprising

- means for acting as a server when connected to other nodes and other possible data transfer channels, such as those connected to an external communications network, and further to a nearly on-line data base, such as a magnetic or optical mass storage, or alternatively, to accessory equipment serving data transfer between the nodes, and still further, to a central management unit,

- means for delivering data contained in the server to at least one dedicated immediate user,

- data storage means (2) formed by a magnetic, optical or solid-state memory associated with a controller, serving both data connections for the dedicated users (5) and at least one input (7) and at least two outputs (6) connected to a routing device (3) and having sufficient capacity to store the data of at least one continuously playable video and/or audio sequence of an average length,

- a routing device (3) capable of routing data transfers from other nodes or data transfer devices or from said data storage means (2) of the node itself to other nodes or data transfer units or to the data storage means (2) of the node itself, in which data transfer a plurality of data streams can be carried in a single physical data transfer channel, and

- a control device (4) suitable for controlling the functions of the data storage means (2) and the routing device (3), whereby

- said control device (4) is capable of receiving control data from the dedicated users of the node (1) and from the possible central management system through an appropriate connection (10),

characterized by

- having a capability of distributing data from its data storage means (2) simultaneously both to the dedicated users (1) of the node and to at least two other nodes or data transfer units, and

- having a capability of routing data transfer from the inputs to the outputs whenever it does not need access to the contents of the transferred data, and

- having a capability of storing a file into the data storage (2) whenever the node has an internal need for gaining access to the transferred data, wherein

- the node is capable of starting the delivery of a file to other nodes and data transfer units and the dedicated users even if the entire content of the file has not yet been copied into the node (2).

2. The node as defined in claim 1, characterized in that the data connections (5) serving the dedicated users of the node pass through said routing device (3).

3. The node as defined in claim 1 or 2, characterized by

- a storage space reserved for the user-specific elastic ring buffers (13) in the data storage means (2) of the node,

- an optional plurality of elastic ring buffers serving user request for files needing different data transfer rates,

- an arrangement permitting each elastic ring buffer (13) to store the data coming from other nodes or accessory equipment under a request of the dedicated user of the node,

- each elastic ring buffer (13) having its own connection in the routing device to obtain the incoming data,

- in addition to the data connection (5) of its dedicated user, said node being capable of delivering a file from its elastic ring buffer (13) via the output of its data storage (7) and the routing device (3) to at least two outgoing data transfer links (9) from the node with the provision that, at the beginning of said intended file transfer, the start point of the file exists in said elastic ring buffer (13), and

- said data transfer connection having the capability of reducing the data write speed so as to prevent the write pointer (15) from passing-by any of the read pointers (14).

4. The node as defined in any of claims 1 - 3, characterized in that

- the control data of the node is transferred in the same incoming (8) and outgoing (9) channels as are used for file transfer,

- the incoming control data is separated from the signal of the incoming data transfer channels by means of a control data separation device (11) connected to said routing device (3), and

- the outgoing control data is connected onto the outgoing data transfer channels by means of connection device (12) operating separately or connected to said routing device (3).

5. The node as defined in any of claims 1 - 4, characterized in that said node (1) is also capable of receiving information via the data transfer connections of its dedicated users.

6. The node as defined in any of claims 1 - 5, characterized in that, with or in lieu of a user-dedicated connection (5) and/or with or in lieu of a node-specific memory unit (2) and/or a routing device (3), the node has at least one data transfer device connected thereto, such as a connection to an external memory, a data communications network (21) or an information producer or a nearly on-line high-capacity data base (22).

7. A server intended for file distribution and storage, said server being comprised of nodes defined in claims 1 - 6, or combinations thereof and possibly complemented with data transfer, monitoring and control means, characterized in that

- the users of the server receive the information contained in the files via said nodes (1),

- each node (1) communicates with at least three other nodes or data transfer devices, and

- said server has preset a data transfer rate defined separately for each different file type, or alternatively, operates at a minimum data transfer rate sufficiently high for use with all different file types, whereby the data transmission rate employed between the nodes (1) must exceed said preset or minimum data transfer rate(s).

8. The file server as defined in claim 7, characterized in that said filer server has at least one user capable of communicating with a number of nodes (1) simultaneously.

9. The file server as defined in claim 7 or 8, characterized in that data transfer between any given pair of nodes can be performed over more than one route.

10. The file server as defined in any of foregoing claims 7 - 9, characterized by

- data transfer devices (22) incorporated in said file server, said devices being capable of distributing files of identical content, and

- having said data transfer devices with the capability of distributing files of identical content located in a decentralized manner in the separate nodes (1) of the file server so that separate routes can be established from said nodes to the data communications nodes defined in claim 6.

11. The file server as defined in any of foregoing claims 7 - 10, characterized by

- said file server being formed by structural entities of different levels so that the structures of a lower level are interconnected forming a similar topology as the interconnected structures of an upper level,

- a lowest-level structure formed by an entity of at least four said nodes (1), and

- data transfer taking place between the different structures via data transfer channels (8, 9) of said nodes (1).

12. The file served as defined in claim 11, characterized in that the file server topology is an N-dimensional hypercube formed by copying an (N-1)-dimensional hypercube, whereby a 0-dimensional hypercube is defined to contain one node, and then establishing data connections via the data transfer channels (8, 9) of the nodes between the corresponding nodes of the (N-1)-dimensional hypercubes thus formed.

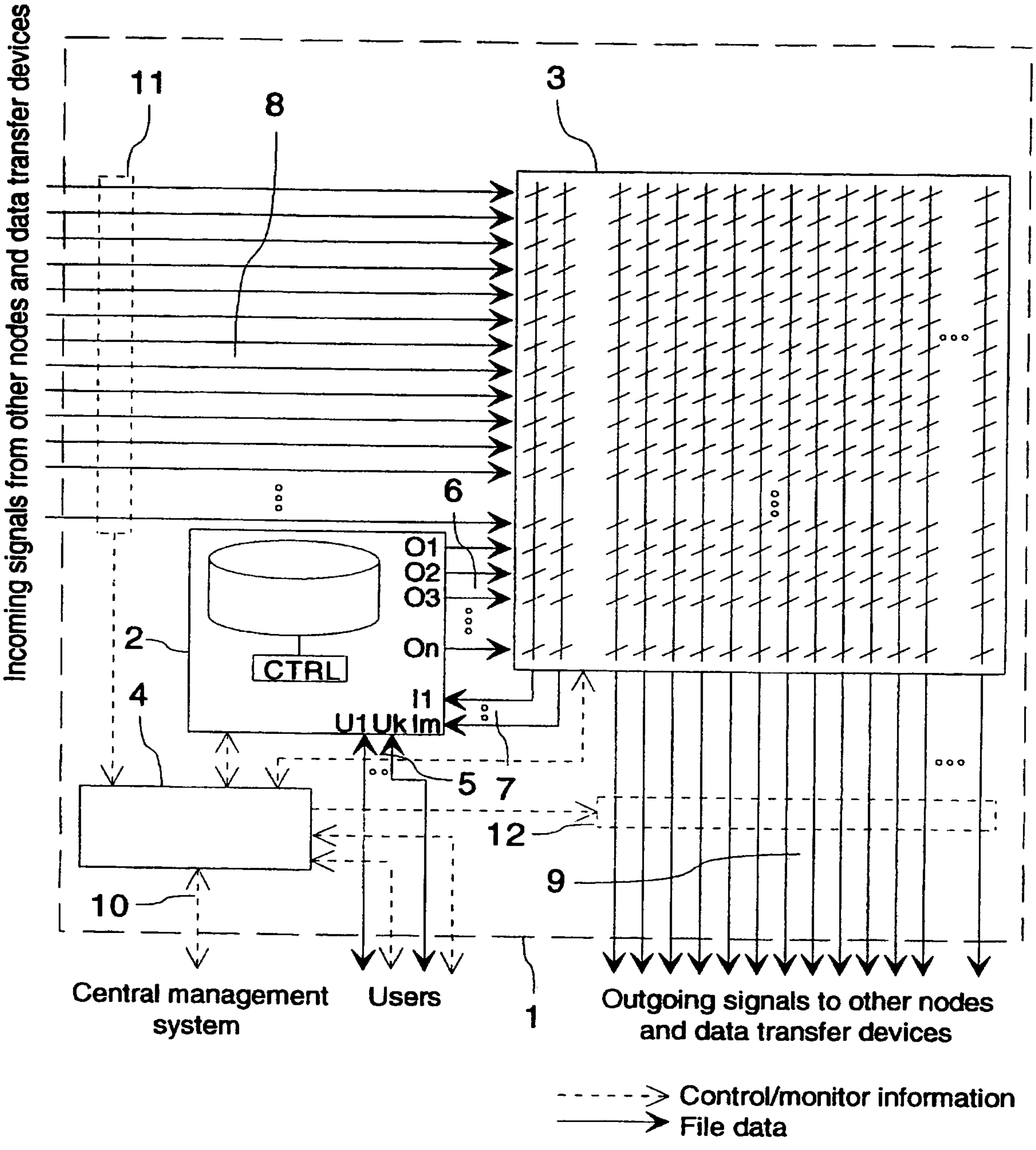

Incoming signals from other nodes and data transfer devices
11
8
3
6
O1
O2
O3
On
CTRL
2
4
I1
U1
Uk
Im
5
7
12
9
10
Central management system
Users
1
Outgoing signals to other nodes and data transfer devices
Control/monitor information
File data

FIG.1

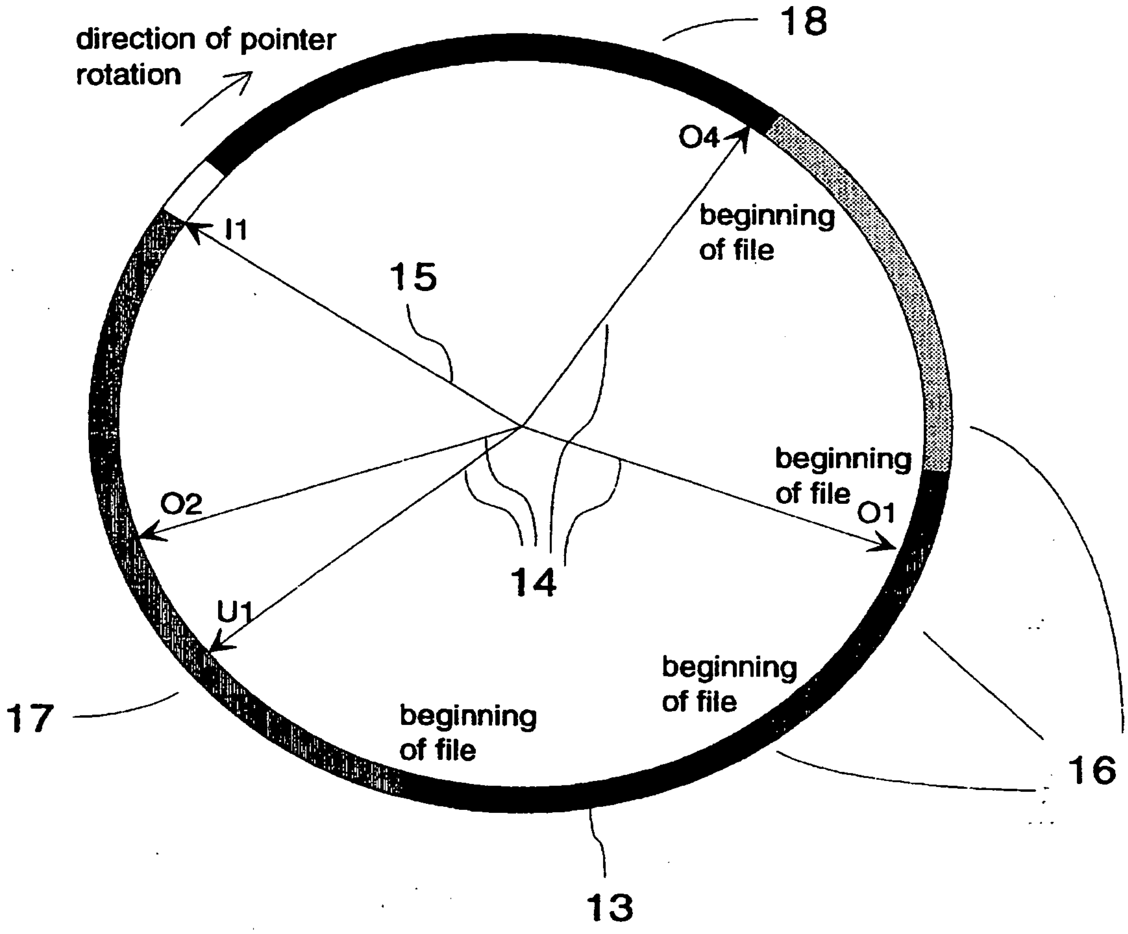

direction of pointer rotation
18
O4
I1
beginning of file
15
beginning of file
O1
O2
14
U1
beginning of file
beginning of file
17
16
13

FIG.2

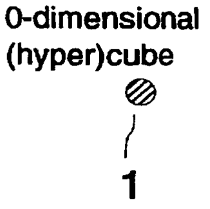


FIG.3A

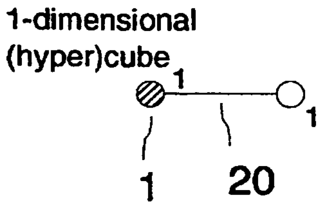


FIG.3B

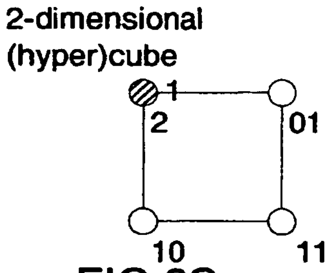


FIG.3C

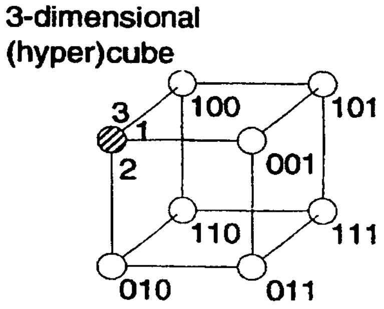


FIG.3D

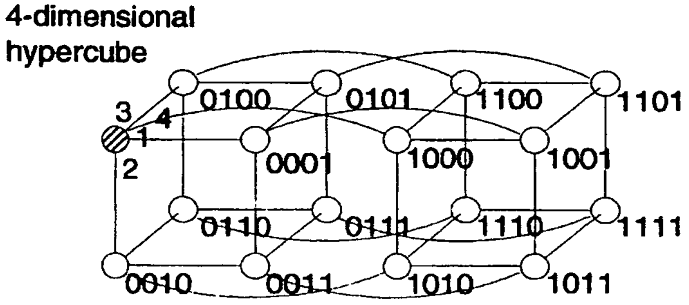


FIG.3E

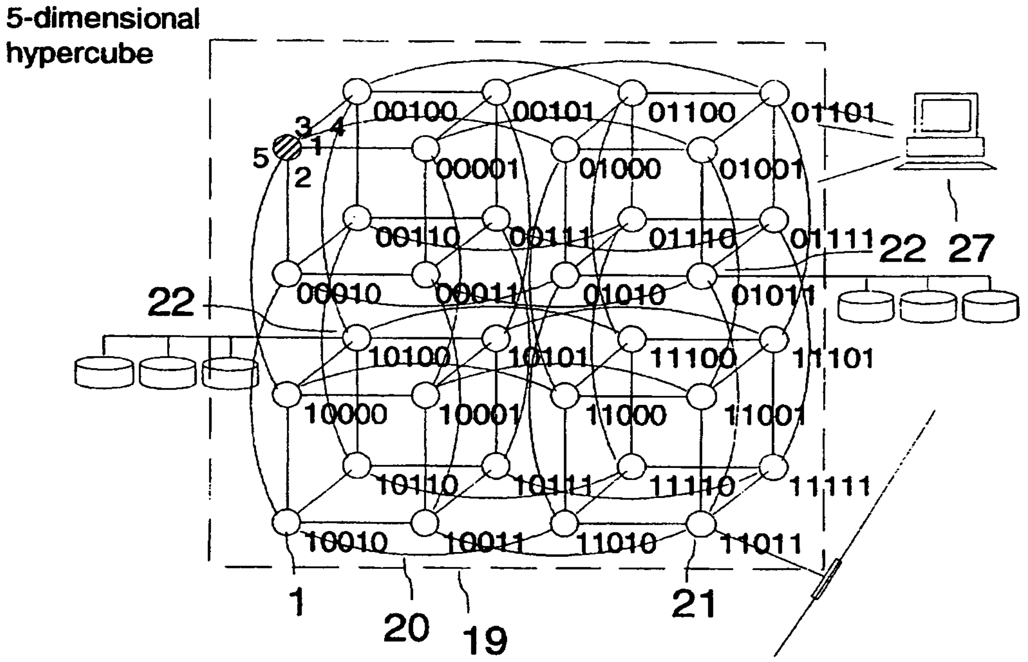


FIG.3F

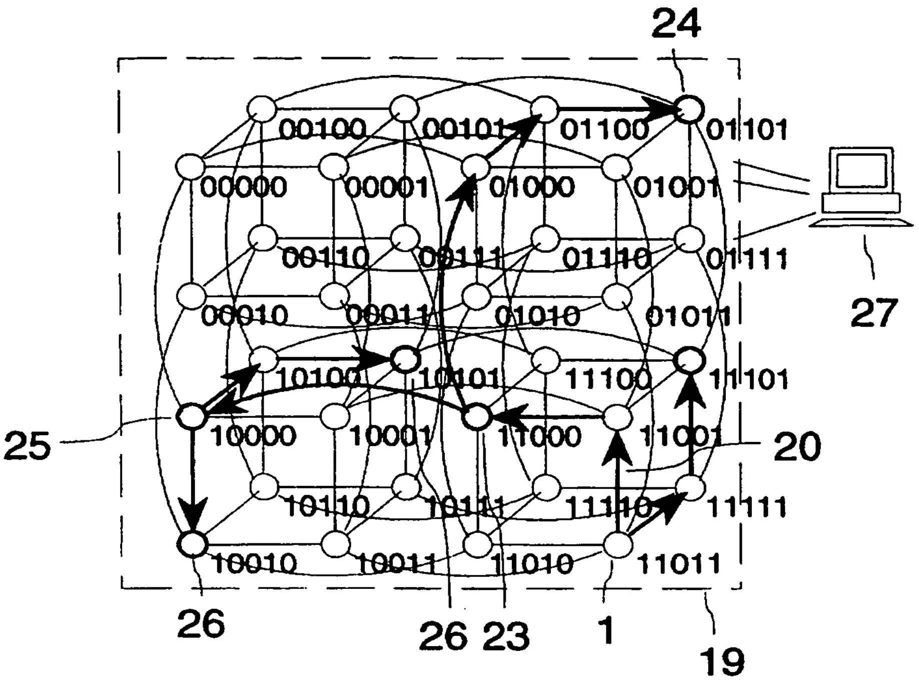

24
00100
00101
01100
01101
00000
00001
01000
01001
00110
00111
01110
01111
00010
00011
01010
01011
27
10100
10101
11100
11101
25
10000
10001
11000
11001
20
10110
10111
11110
11111
10010
10011
11010
11011
26
26
23
1
19

FIG.4